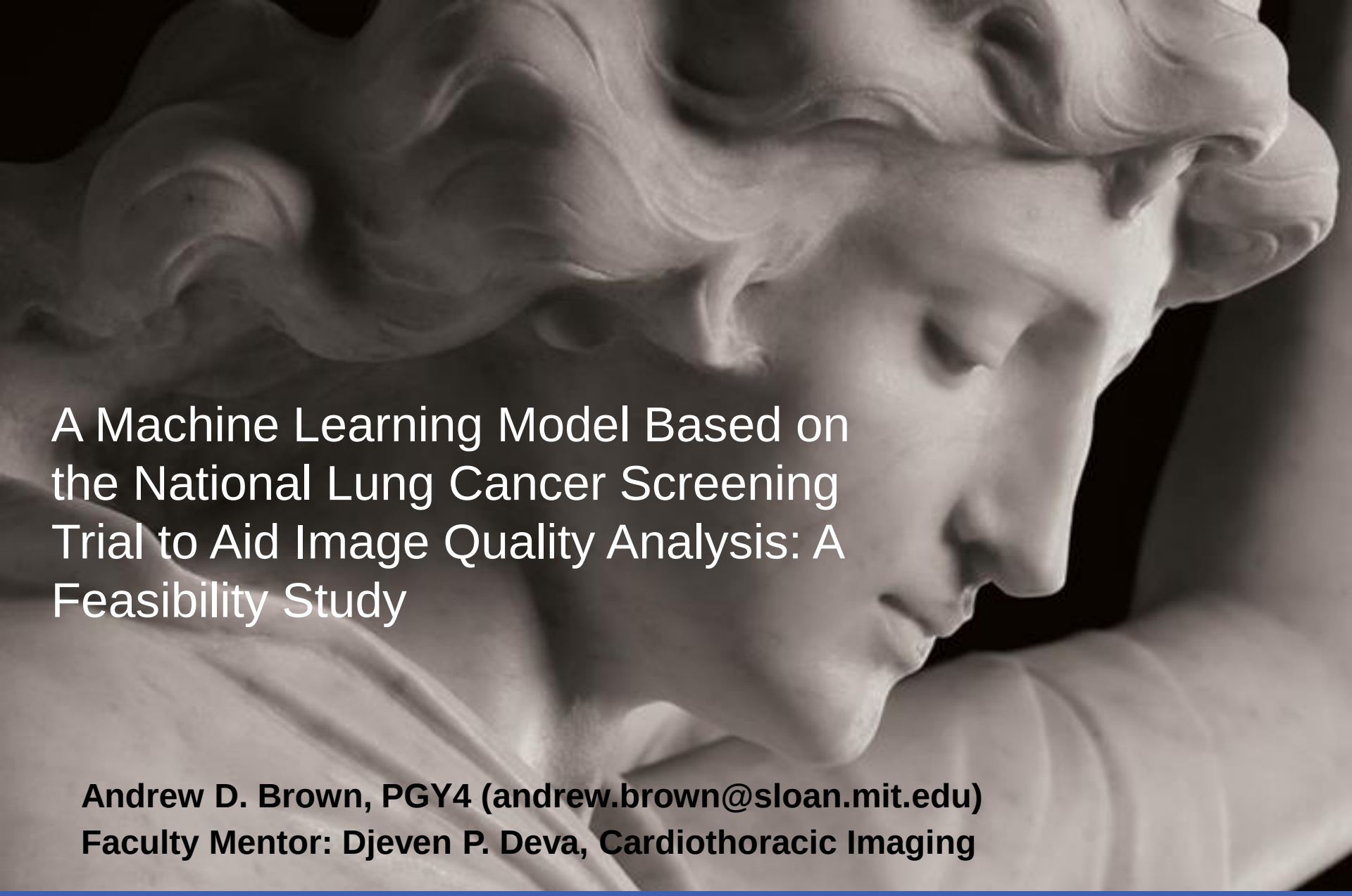


A Machine Learning Model Based on the National Lung Cancer Screening Trial to Aid Image Quality Analysis: A Feasibility Study

Andrew D. Brown, PGY4 (andrew.brown@sloan.mit.edu)
Faculty Mentor: Djeven P. Deva, Cardiothoracic Imaging

UNIVERSITY OF TORONTO
FACULTY OF MEDICINE

St. Michael's
Inspired Care. Inspiring Science.

Disclosures

None

Background

Image quality standards for providers of CT, MR, and nuclear medicine imaging are an important mechanism for maintaining high quality care for patients receiving diagnostic imaging

However, the costly and time-consuming nature of manual image analysis present challenges for many quality assurance programs

For example, the American College of Radiology (ACR) CT Accreditation program evaluates facilities on a number of categories including technical parameters, anatomic coverage, artifacts (such as motion and metallic streaks), exam identification and examination protocols[1]

The ACR CT Accreditation program takes **4 to 6 months**[2]

Background

Such programs have several limitations[3,4], including:

- *Infrequent assessments*
  - Gaps of 3 to 4 years between accreditation cycles
- *Binary feedback*
  - Simply a pass-or-fail grading system
- *"Cherry-picking"*
  - Allowing imaging providers to select their best images for evaluation

Automated image analysis may provide an opportunity to conduct reviews more frequently, provide dynamic measures of quality and examine larger, more representative samples of images than the current process allows

Research question

Can machine learning techniques be used to identify image quality deficiencies, specifically motion artifacts, in CT examinations?

In this work, we chose to focus on motion artifacts as they are currently evaluated as part of the ACR CT accreditation process and represent an area we believed machines could learn from experienced readers

Methods

Approaches for developing machine learning models often involve identifying a dataset which reflects the real world problem of interest[5]

For this study we examined CT scans performed as part of the National Lung Screening Trial (NLST)[6]

NLST was a randomized trial conducted to compare low-dose CT scans with chest X-rays as a means of screening for lung cancer

The presence or absence of motion artifacts was identified as part of the quality control procedure during the study

Methods - Data

From the National Cancer Institute we requested all examinations from participants in the low-dose CT arm of the NLST that had been identified as having image quality deficiencies as well as a random sample of examinations without image quality deficiencies

Incomplete screening examinations and those with missing or corrupted images were excluded from our study

Our complete dataset consisted of:

- 911 **no-motion** low-dose CT chest examinations
- 466 **motion** degraded low-dose CT chest examinations

A lung window (level 600 HU and window 1500 HU) was applied to each CT examination

Methods - Image analysis

A general purpose image sharpness estimator, the Perceptual Sharpness Index (PSI)[7], was applied to each image in the CT examination to measures the sharpness or blurriness of images

Sharpness estimation examines both the edges and contrast within an image for blurring characteristics, such as the smearing of adjacent structures

The PSI is an index between 0 and 1, with higher scores signifying sharper images

Methods - Image analysis

We measured the sharpness of every image in each low-dose CT screening exam in our dataset to create a distribution of PSI scores for each examination

These distributions were summarized by 10 different statistics such as mean, min/max, range and skew

We then applied a machine learning algorithm, Multivariate adaptive regression splines (MARS)[8], to identify any pattern in these statistics that might predict image quality deficiencies

Methods - Baseline

Often in machine learning, proposed statistical approaches are compared to a baseline - a simple model which aims to explain the relationship between the dependant and independent variables[5]

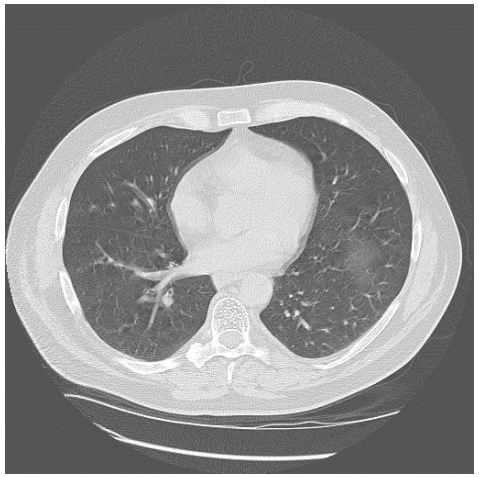
In our study we used a combination of demographic data and technical parameters of the CT study as independent variables to predict motion degradation as identified by a radiologist

Demographic data and technical parameters were chosen as the inputs in the baseline model as these are thought to contribute to deficiencies

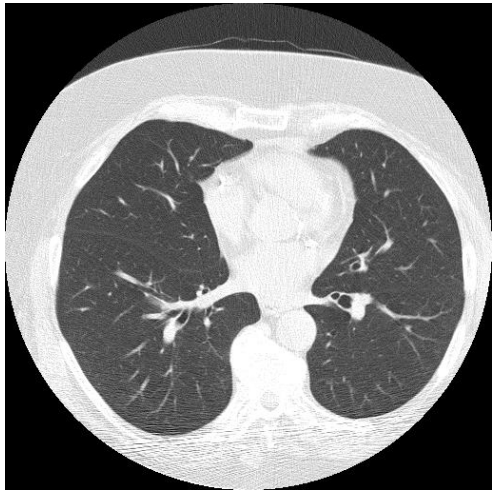
These independent variables were also fed into a MARS model to identify the patterns that might predict image quality deficiencies

Results - Example of PSI measure

Low-dose CT chest from the **motion** subgroup. There is blurring of the heart and pulmonary vasculature consistent with respiratory and cardiac motion. This image had a Perceptual Sharpness Index (PSI) of 0.68272.

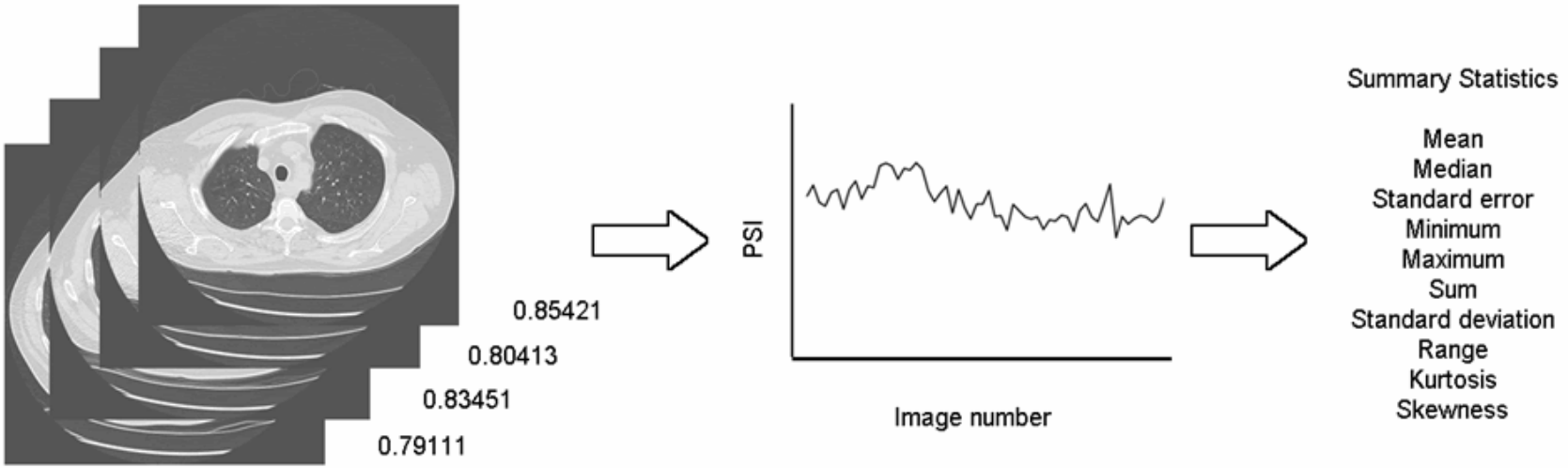


Low-dose CT chest from the **no-motion** subgroup. This image appears sharper than the motion degraded image and had a Perceptual Sharpness Index (PSI) of 0.89406.



Results - Sharpness estimation process

Each image in a CT examination was analyzed and given a PSI score. The constellation of PSI scores creates a distribution, which was converted to a collection of summary statistics for each examination.



Results - Model evaluation

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Baseline	65 (60-70)	17 (10-25)	87 (83-91)
PSI+MARS	94 (91-96)	100 (95-100)	91 (87-94)

* 95% confidence interval

Results - Relative importance of the independent variables

To identify the three most important variables in the model we employed a form of the residual sum of squares to rank the model predictors on a 0-100 scale

Baseline Model		PSI+MARS Model	
Variable	**Importance score**	**Variable**	**Importance score**
Weight	100	Mean	100
Manufacturer A	57	Max	84
Manufacturer B	43	Min	39

Discussion

In the PSI+MARS model, we use the PSI summary statistics as inputs in a MARS model. Our PSI+MARS model outperformed the baseline on every performance measure.

The most important variable in the PSI+MARS model was the mean PSI of the CT examination. This makes intuitive sense, as the mean is a measure of the central tendency of a distribution and summarizes the overall sharpness of the examination. Maximum PSI in a examination was the second most important variable and was found to be relatively more important than the minimum PSI score.

An interesting area for further research would be to examine whether radiologists share our model's preference for maximum sharpness as the most important predictor of image quality.

Discussion

The baseline model consisted of patient demographic information and technical parameters from the CT scan unit.

The most important demographic variable for predicting motion artifacts was weight. This is consistent with other research, which suggests that patient factors such as age, body habitus and mobility may influence exam complexity and in turn image quality[10]. We found that age and gender were not significant factors.

The most important technical parameter was the scanner manufacturer. We hypothesize this was related to the age of the scanners. However, this is difficult to interpret with the limited manufacturer information within the NSLT dataset.

Limitations

Our gold standard for image assessment was the ability of radiologists in the NLST to accurately identify motion artifacts. In the NLST all radiologists received training in the assessment of suboptimal image quality. While this may have reduced any potential variability between radiologists, it cannot be eliminated completely.

This preliminary study explored the feasibility of detecting motion artifacts however we did not assess the degree to which the presence of motion impaired the reader's ability to screen for lung cancer. This is important but beyond the scope of this work.

Further development of tools such as our model will be required to validate the accuracy, effectiveness, and generalizability of artificial intelligence in radiology quality measurement.

Summary

In this study we have combined two data driven approaches, PSI and MARS, to automate image quality analysis

Our model was able to effectively discriminate between motion-degraded and non-motion degraded low-dose CT examinations with a high degree of accuracy, outperforming the baseline model

Automated image analysis may allow for the analysis of more images, while improving the speed, reliability and cost of quality assurance

References

1. American College of Radiology. CT Accreditation Program. Testing Instructions. Available at: http://www.acraccreditation.org/~/media/ACRAccreditation/Documents/CT/CT-Accreditation-Testing-Instructions.pdf. Accessed March 26, 2017.
2. American College of Radiology. CT Accreditation Program FAQ. Available at: http://www.acraccreditation.org/How-To/CT-Accreditation-FAQ. Accessed March 26, 2017.
3. Reiner BI. Creating Accountability in Image Quality Analysis. Part 2: Medical Imaging Accreditation. J Digit Imaging 2013; 26:371-374.
4. Garvey CJ, Cook JV, Wiltsher C, Whitley S. Radiology accreditation—towards a safer quality service. Clin Radiol 2009; 64:853-856.
5. Kuhn M, Johnson K. Applied predictive modeling. New York, NY: Springer, 2013:61-80.
6. Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011; 365:395-409.
7. Feichtenhofer C, Fassold H, Schallauer P. A perceptual image sharpness metric based on local edge gradient analysis. IEEE Signal Process Lett 2013; 20:379-382.
8. Friedman JH. Multivariate adaptive regression splines. Ann Stat 1991; 1-67.
9. Friedman JH, Roosen CB. An introduction to multivariate adaptive regression splines. Stat Methods in Med Res 1995; 4:197-217
10. Reiner BI. Creating Accountability in Image Quality Analysis. Part 4: Quality Analytics. J Digit Imaging 2013; 26:825–829